Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence

Year: 2023 | Citations: 948 | Authors: Vikas Hassija, V. Chamola, Atmesh Mahapatra, Abhinandan Singal, Divyansh Goel

---

Abstract

Recent years have seen a tremendous growth in Artificial Intelligence (AI)-based methodological development in a broad range of domains. In this rapidly evolving field, large number of methods are being reported using machine learning (ML) and Deep Learning (DL) models. Majority of these models are inherently complex and lacks explanations of the decision making process causing these models to be termed as 'Black-Box'. One of the major bottlenecks to adopt such models in mission-critical application domains, such as banking, e-commerce, healthcare, and public services and safety, is the difficulty in interpreting them. Due to the rapid proleferation of these AI models, explaining their learning and decision making process are getting harder which require transparency and easy predictability. Aiming to collate the current state-of-the-art in interpreting the black-box models, this study provides a comprehensive analysis of the explainable AI (XAI) models. To reduce false negative and false positive outcomes of these back-box models, finding flaws in them is still difficult and inefficient. In this paper, the development of XAI is reviewed meticulously through careful selection and analysis of the current state-of-the-art of XAI research. It also provides a comprehensive and in-depth evaluation of the XAI frameworks and their efficacy to serve as a starting point of XAI for applied and theoretical researchers. Towards the end, it highlights emerging and critical issues pertaining to XAI research to showcase major, model-specific trends for better explanation, enhanced transparency, and improved prediction accuracy.